CHAL: Causal-Guided Hierarchical Anomaly-Aware Learning for Moving Infrared Small Target Detection (Supplementary Material)

A. Overall

In this supplementary material, we summarize the differences between our proposed CHAL framework and existing works in **Section B**. In **Section C**, we provide more details about our CHAL, including theoretical analysis, pseudo code, and the relations of each component. Then, in **Section D**, the details of three public infrared datasets $i.e.$, DAUB-H, NUDT-MIRSDT and IRDST-R, and a visible-light remote sensing dataset, $i.e.$, RsCarData, are presented. In **Section E**, we provide more experimental results. These include: (1) quantitative comparisons; (2) inference cost comparisons; (3) PR curve comparisons; and (4) the ablation studies to analyze: different background modeling methods; effects of cosine similarity; effects of Spatio-Temporal Neural Fields; effects of Hierarchical Anomaly-aware Learning; effects of Causal Relation Guiding; effects of Anomaly Strength and Type; and hyperparameter study of CHAL. These additional experimental results consistently demonstrate the effectiveness and superiority of our proposed CHAL. In **Section F**, we present failure case analysis and further discussion. Finally, in **Section G**, more visualization comparisons are provided.

B. Differences with Existing Works

For clarity and intuitiveness, we summarize the typical differences between our CHAL and existing works, as shown in Figure 1. From figure, three typical differences emerge.

(I) Learning Paradigm: Almost all previous methods for infrared small targets, $e.g.$, ISNet [27], Tridos [3] and DTUM [7], are based on target-centered learning, trying to learn the weak features of infrared targets directly from cluttered backgrounds. In contrast, our CHAL develops the paradigm of background-centered learning, focusing on dominant backgrounds rather than ambiguous targets.

(II) Problem Definition: Previous works often treat moving infrared small target detection (MISTD) as a feature matching or classifying problem, resulting in a fundamental contradiction, $i.e.$, they typically rely on rich target features, which are precisely lacking in infrared targets. Conversely, our CHAL reformulates MISTD as an anomaly discovery task, regarding targets as the anomalies that deviate from the spatio-temporal evolution patterns of backgrounds.

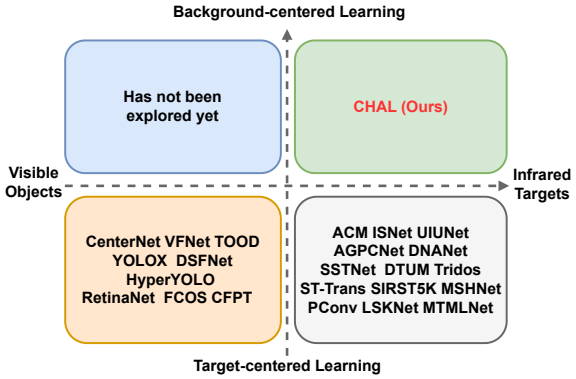

Figure 1. Typical differences of our CHAL and existing works.

(III) Application Scenarios: At present, the proposed background-centered learning has not been explored adequately. Our CHAL is specifically designed for moving infrared small targets in more challenging scenarios than those involving visible objects. We provide the first in-depth investigation to address the gap in background-centered learning for infrared small targets. Moreover, even in visible-light scenarios, it still exhibits obvious adaptivity.

C. Details of Proposed CHAL

C.1. Theoretical Analysis

Causal Relation Guiding The goal of our designed CRG is to block the confounding path $\boldsymbol{F} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{Y}$ and estimate the true causality $P(\boldsymbol{Y}|\text{do}(\boldsymbol{F}))$. The standard backdoor adjustment formula is as follows:

$$P(\boldsymbol{Y}|\text{do}(\boldsymbol{F})) = \sum_{z} P(\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{Z}=z)P(\boldsymbol{Z}=z), \quad (1)$$

where $do\ (\cdot)$ signifies an intervention to remove the influence of $\boldsymbol{Z}$, $P(\boldsymbol{Z}=z)$ is the probability density function of $\boldsymbol{Z}$. However, because $\boldsymbol{Z}$ in MISTD is continuous and dynamic, direct computation is infeasible. As such, our CRG

employs differentiable operations (Eq. (10) and (11) in Section 3.4) for approximation. For formal proof this, we introduce the following assumptions:

Assumption 1 (Causal Graph). Variable relationships are described by a directed acyclic graph, and $\boldsymbol{Z}$ satisfies the backdoor criterion.

Assumption 2 (Ignorability). Given background confounders $\boldsymbol{Z}$, the intervention $\text{do}(\boldsymbol{F})$ and $\boldsymbol{Y}$ are conditionally independent, *i.e.*, $\boldsymbol{Y} \perp\!\!\!\perp \text{do}(\boldsymbol{F}) \mid \boldsymbol{Z}$.

Assumption 3 (Smoothness): $P(\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{Z} = z)$ and $P(\boldsymbol{Z} = z)$ are Lipschitz continuous with z.

Assumption 4 (Consistency): The normalized anomaly score $\hat{\boldsymbol{Z}} = f_{nor}(\boldsymbol{\mathcal{A}_f})$ is a consistent proxy for $\boldsymbol{Z}$, *i.e.*, $\hat{\boldsymbol{Z}} \to \boldsymbol{Z}$ in probability as the number of samples increases.

Under these assumptions, we could further prove three core propositions about our CRG, as follows:

Proposition 1 (Unbiasedness of Back-door Adjustment): Under Assumptions 1 and 2, the back-door adjustment formula can provide an unbiased estimation of the true causal effect, which is formulated as follows:

$$P(\boldsymbol{Y}|\text{do}(\boldsymbol{F})) = \int_z P(\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{Z} = z)P(\boldsymbol{Z} = z)dz. \quad (2)$$

Proof. According to the back-door criterion [12], if $\boldsymbol{Z}$ blocks all back-door paths from $\boldsymbol{F}$ to $\boldsymbol{Y}$, the interventional distribution is identifiable by the integral formula above. Under Assumption 1, $\boldsymbol{Z}$ satisfies the back-door criterion. Assumption 2 ensures that conditional ignorability holds. For a continuous $\boldsymbol{Z}$, the integral requires measurability conditions, which are guaranteed by Assumption 3. In the discrete case, the formula degenerates to a summation. Therefore, this estimation is unbiased.

Proposition 2 (Approximate Consistency of CRG): Under Assumptions 3 and 4, the differentiable operation of CRG (Eq. (10) and (11) in main paper) consistently approximates the back-door adjustment formula. Specifically, let $\hat{\boldsymbol{Z}} = f_{nor}(\boldsymbol{\mathcal{A}_f})$, and the anomaly score $\boldsymbol{\mathcal{H}}$ be defined as:

$$\boldsymbol{\mathcal{H}} = \begin{cases} \eta_e \cdot \hat{\boldsymbol{Z}}, \text{ if } \hat{\boldsymbol{Z}} > \tau, \\ \eta_s \cdot \hat{\boldsymbol{Z}}, \text{ otherwise}, \end{cases} \quad (3)$$

where $\eta_e > 1$, $\eta_s < 1$, and τ is a threshold. Then the adjusted feature $\boldsymbol{F_f}$ (generated by Eq. (11)) satisfies:

$$\lim_{\hat{\boldsymbol{Z}} \to \boldsymbol{Z}} P(\boldsymbol{Y}|\boldsymbol{F_f}) = P(\boldsymbol{Y} \mid \text{do}(\boldsymbol{F})) \quad (4)$$

Proof. The CRG mechanism approximates a stratification strategy by partitioning the continuous $\boldsymbol{Z}$ into two strata (high anomaly and low anomaly), achieving the approximation via feature intervention. As $\hat{\boldsymbol{Z}} \to \boldsymbol{Z}$ by Assumption 4, $\boldsymbol{\mathcal{H}}$ approximates a weighted indicator function:

$$\boldsymbol{\mathcal{H}} \approx \mathbb{I}(\boldsymbol{Z} > \tau) \cdot \eta_e \boldsymbol{Z} + \mathbb{I}(\boldsymbol{Z} \leq \tau) \cdot \eta_s \boldsymbol{Z} \quad (5)$$

This effectively re-weights the distribution $P(\boldsymbol{Z} = z)$. The operation $\hat{\boldsymbol{F}}_{\boldsymbol{t}} \odot (1 + \boldsymbol{\mathcal{H}})$ in Eq. (11) introduces an importance weight $\boldsymbol{\mathcal{W}}(z) = 1 + \boldsymbol{\mathcal{H}}(z)$, that is:

$$\boldsymbol{F_f} \approx \int_z \boldsymbol{F_t} \cdot \boldsymbol{\mathcal{W}}(z) P(\boldsymbol{Z} = z)dz. \quad (6)$$

Under Assumption 3, the continuity of $f_{ref}(\cdot)$ ensures:

$$P(\boldsymbol{Y}|\boldsymbol{F_f}) = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{F_f}] \approx \int_z \mathbb{E}[\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{Z} = z]\boldsymbol{\mathcal{W}}(z)p(z)dz. \quad (7)$$

By the Law of Large Numbers and Assumption 4, as $\hat{\boldsymbol{Z}} \to \boldsymbol{Z}$, we have $\boldsymbol{\mathcal{W}}(z) \to 1$. As such,

$$P(\boldsymbol{Y}|\boldsymbol{F_f}) \to \int_z P(\boldsymbol{Y}|\boldsymbol{F}, \boldsymbol{Z} = z)p(z)dz = P(\boldsymbol{Y}|\text{do}(\boldsymbol{F})), \quad (8)$$

the approximation error is bounded by Lipschitz continuity.

Proposition 3 (Optimality of the Loss Function): Minimizing the total loss $\mathcal{L}_{total}$ drives the model parameters θ to estimate the true causal effect, as follows:

$$\arg\min_{\theta} \mathcal{L}_{total} \implies P(\boldsymbol{Y}|\boldsymbol{F_f}) \to P(\boldsymbol{Y} \mid \text{do}(\boldsymbol{F})), \quad (9)$$

where θ are the model parameters, $\mathcal{L}_{total} = \mathcal{L}_{obj} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{cls}$, as shown in Eq. (12) in the main paper.

Proof. From **Proposition 2**, $\boldsymbol{F_f}$ is an approximately unbiased feature representation. The loss $\mathcal{L}_{total}$ is computed on $\boldsymbol{F_f}$ and implicitly supervises all upstream components (SNF, HAL, and CRG). Consider the expected risk:

$$\mathcal{L}_{total} = \mathbb{E}_{(I,\boldsymbol{Y})\sim\mathcal{D}}[\ell(\hat{\boldsymbol{Y}}(\boldsymbol{F_f}), \boldsymbol{Y})], \quad (10)$$

where ℓ is the composite loss (using sigmoid focal loss for classification and NWD loss for localization). The focal loss is a consistent loss function [8], and its minimization ensures that the prediction $\hat{\boldsymbol{Y}}(\boldsymbol{F_f})$ converges to the expectation $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{F_f}]$. Combined with Proposition 2, we have:

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{F_f}] \to P(\boldsymbol{Y} \mid \text{do}(\boldsymbol{F})) \quad (11)$$

The gradient back-propagates through $\boldsymbol{F_f}$ to penalize residual deviations, ensuring overall convergence. By proxy, minimizing $\mathcal{L}_{total}$ is equivalent to minimizing the KL divergence between the estimated distribution $P(\boldsymbol{Y}|\boldsymbol{F_f})$ and the true causal effect $P(\boldsymbol{Y}|\text{do}(\boldsymbol{F}))$.

In this way, we have rigorously proved that the adjustment term $\boldsymbol{\mathcal{H}}$ used in our CRG is not a heuristic design. Instead, it is a differentiable approximation of ideal back-door adjustment. By setting η_e, η_s as learnable parameters, our CRG could automatically learn the optimal parameters in an end-to-end manner, with the implicit supervision of $\mathcal{L}_{total}$. This could effectively block spurious correlations and enhance true causality $\boldsymbol{T} \to \boldsymbol{F} \to \boldsymbol{Y}$.

Algorithm 1 Our Proposed CHAL Framework

Require: Infrared video sequences $\boldsymbol{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots \boldsymbol{I}_t\}$, Loss weights λ_1, λ_2, and causal anomaly threshold τ.
Ensure: Final detection results $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$ for $\boldsymbol{I}_t$, total loss $\mathcal{L}_{total}$.

Initialize $\boldsymbol{F}$ and $\boldsymbol{F_C}$ as empty list
for k from 1 to t **do** ▷ extract multi-scale features
  $\boldsymbol{F_k}$ = CSPDarknet ($\boldsymbol{I}_k$), $\boldsymbol{F} \leftarrow \boldsymbol{F_k}$
end for
Obtain multi-frame features $\boldsymbol{F_C}$ by early feature fusion.
// Spatio-Temporal Neural Fields
Construct spatio-temporal volume $\boldsymbol{V}$ by stacking $\boldsymbol{F_C}$.
$\boldsymbol{Z_s} = \boldsymbol{\psi_s}(\boldsymbol{V})$ ▷ appearance scene encoding
$\boldsymbol{Z_d} = \boldsymbol{\psi_d}(\boldsymbol{V})$ ▷ dynamic scene encoding
Fuse $\boldsymbol{Z_s}$ and $\boldsymbol{Z_d}$ to obtain overall scene encoding $\boldsymbol{Z}$
Generate query signals $\boldsymbol{Q}$ by the position encoder $\boldsymbol{\mathcal{P}}$, 3D coordinate grid $\boldsymbol{\mathcal{G}}$ and Fourier feature mapping
$\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{F}_\theta}(\boldsymbol{Z}, \boldsymbol{Q})$ ▷ neural field decoding for background
// Hierarchical Anomaly-Aware Learning
for k from 1 to t **do** ▷ compute appearance anomaly
  $\boldsymbol{\mathcal{A}_a^k} = \Delta_\theta(1 - f_{cos}(\boldsymbol{F}_k, \boldsymbol{\mathcal{B}}_k))$, $\boldsymbol{\mathcal{A}_a} \leftarrow \boldsymbol{\mathcal{A}_a^k}$
end for
Verify $\boldsymbol{\mathcal{A}_a}$ by motion consistency to obtain $\boldsymbol{\mathcal{A}_c}$
Reconstruct anomaly to obtain $\boldsymbol{\mathcal{A}_r}$ by a decoder $\Phi_r(\cdot)$
$\boldsymbol{\mathcal{A}_f} = \sum_{k=1}^{t} \omega_k \cdot \boldsymbol{\mathcal{A}_r^k}$ ▷ temporal weight learning
// Causal Relation Guiding
Normalize $\boldsymbol{\mathcal{A}_f}$ to be a proxy for confounders $\boldsymbol{Z}$
Perform causal intervention with anomaly scores $\boldsymbol{\mathcal{H}}$
$\boldsymbol{F_f} \leftarrow f_{ref}(\boldsymbol{F_g})$ ▷ refinement for adjusted $\boldsymbol{F_g}$
Feed into a detection head to infer $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$ or calculate $\mathcal{L}_{total}$
return $\hat{\boldsymbol{Y}}_{\boldsymbol{t}}$ or training loss $\mathcal{L}_{total}$

C.2. Pseudo Code

To better illustrate the implementation of the proposed CHAL, we provide the detailed computation process, as shown in Algorithm 1. From the Algorithm, it is obvious to observe the data flow and synergistic interactions among the three core components of our CHAL, *i.e.*, SNF, HAL and CRG. They are not a simple stacking but a carefully designed data processing chain.

First, SNF is the foundation of HAL. It exploits the continuous modeling capability of neural fields to precisely reconstruct the expected normal background $\boldsymbol{\mathcal{B}}$. $\boldsymbol{\mathcal{B}}$ is fed into the HAL along with the multi-frame feature $\boldsymbol{F_C}$ to calculate the appearance anomaly for each frame.

Second, HAL provides the causal intervention proxy for CRG. It acts as an anomaly purifier. HAL first captures the appearance anomaly $\boldsymbol{\mathcal{A}_a}$ with numerous suspicious anomalies caused by background confounders $\boldsymbol{Z}$. Then, verify them by motion consistency. HAL could extract the clean signal $\boldsymbol{\mathcal{A}_f}$, containing only high-confidence anomalies, which serves as a learnable proxy for the unobservable confounding factor $\boldsymbol{Z}$. In this way, CRG could precisely perform backdoor adjustment for observed features $\hat{\boldsymbol{F}}_{\boldsymbol{t}}$.

Finally, the output of CRG is the unconfounded feature $\boldsymbol{F_f}$. $\boldsymbol{F_f}$ is fed into a detection head to obtain final results. This is the ultimate goal of our entire framework. We do not want the detector to employ the feature $\hat{\boldsymbol{F}}_{\boldsymbol{t}}$ contaminated by $\boldsymbol{Z}$, but rather the clean feature $\boldsymbol{F_f}$ that could reflect the true causal effect $P(\boldsymbol{Y}|do(\boldsymbol{F}))$.

Therefore, our CHAL is a logically rigorous and collaborative framework. It could perfectly realize a new causal-guided anomaly discovery paradigm by the data flow implicitly driven by $\mathcal{L}_{total}$, *i.e.*, SNF models normal backgrounds → HAL finds anomalies and refines them to be a causal proxy → CRG performs backdoor adjustment to obtain the final unconfounded feature for detection.

D. Datasets and Evaluation Metrics

We evaluate our proposed CHAL on three public moving infrared small target detection datasets: DAUB-H [4], NUDT-MIRSDT [7] and IRDST-R [14], and a visible-light remote sensing dataset, *i.e.*, RsCarData [20].

For DAUB-H, the training set comprises 8 videos with 8,596 frames, and the test set includes 9 videos with 5,181 frames. For NUDT-MIRSDT, we generate bounding box annotations based on the masks. The training set contains 63 videos with 6,300 frames, and the test set contains 47 videos with 4,700 frames. For IRDST-R, the training set consists of 23 videos with 11,768 frames, and the test set consists of 7 videos with 7,339 frames. For RsCarData, the training set contains 70 videos with 27,420 frames, and the test set contains 7 videos with 2,255 frames. The details of four datasets are shown in Table 1.

To make a fair comparison with other detectors, following SSTNet [1], TMP [30] and Tridos [3], we employ the commonly used metrics in object detection paradigm, *i.e.*, Precision (Pr), Recall (Re), F1 score and the Average Precision with an IoU threshold 0.5 (mAP_{50}). These metrics can be formulated as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{12}$$

where TP, FP and FN represent the number of correct detection targets (True positive), false alarms (False positive) and missed detection targets (False negative), respectively. F1 score is a relatively comprehensive metric, combining both precision and recall. AP denotes the numerical integration of interpolated precision across all recall points. It

Table 1. The details of three infrared and a visible-light small target datasets, *i.e.*, DAUB-H, NUDT-MIRSDT, IRDST-R, and RsCarData.

Dataset	Annotations	Sequential	Classes	Target size	Target type	Number of frames	Scene description
DAUB-H	bounding boxes	✓	1	1 ∼ 10 pixels	drone	13,777	ground background, buildings, air background, plain, suburb, air ground junction background
NUDT-MIRSDT	masks, bounding boxes	✓	1	1 ∼ 80 pixels	bright spot	11,000	ground background, city, plain, suburb, sky sea, land, buildings
IRDST-R	bounding boxes	✓	1	1 ∼ 100 pixels	helicopter, airplane, drone	19,107	ground background, forest, air background, lakes, suburb, clouds, buildings, mountains
RsCarData	bounding boxes	✓	1	1 ∼ 50 pixels	car	29,675	city, road, lakes, suburb, buildings

Table 2. Quantitative comparisons on three public datasets. The best one is marked in **bold**, and the second-best one is underlined. All multi-frame methods utilize five consecutive frames in each iteration. "★" means the method is based on anomaly detection.

	Method	Pub	DAUB-H				NUDT-MIRSDT				IRDST-R			
			mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)	mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)	mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)
Single-frame	ACM [2]	WACV'21	49.84	74.83	67.21	70.81	56.85	70.84	81.22	75.68	57.65	79.34	73.30	76.20
	ISNet [27]	CVPR'22	44.73	56.35	80.99	66.46	68.72	81.62	85.50	83.51	59.27	73.75	81.98	77.65
	UIUNet [19]	TIP'22	49.23	73.72	67.49	70.47	66.63	83.00	81.83	82.41	59.22	76.55	78.47	77.50
	AGPCNet [28]	TAES'23	22.89	37.86	61.67	46.91	68.58	82.38	84.98	83.66	58.01	79.76	73.91	76.72
	DNANet [6]	TIP'23	50.76	71.04	72.01	71.52	68.97	82.87	84.61	83.73	67.43	85.97	79.04	82.36
	RPCANet [17]	WACV'24	52.57	70.74	75.16	72.88	51.54	69.38	75.74	72.42	65.63	81.62	81.16	81.39
	SIRST5K [10]	TGRS'24	40.66	57.63	71.14	63.68	68.91	86.88	80.76	83.71	47.43	71.08	67.03	68.99
	SCTrans [24]	TGRS'24	34.74	54.52	64.89	59.25	72.24	86.42	83.21	84.78	51.75	73.13	72.00	72.56
	MSHNet [9]	CVPR'24	26.47	48.56	55.39	51.75	71.77	86.18	84.74	85.45	60.86	78.54	78.52	75.53
	MLPNet [9]	TGRS'25	49.17	68.00	73.13	70.47	66.79	81.22	83.06	82.13	50.20	75.10	67.34	71.01
	LSKNet [18]	TGRS'25	46.34	56.97	**82.76**	67.48	72.73	85.36	86.07	85.72	65.84	82.66	80.88	81.76
	SAMamba [21]	InfFus'25	20.47	36.89	56.75	44.71	45.98	60.29	77.21	67.71	50.20	75.10	67.34	71.01
	PConv [23]	AAAI'25	48.48	70.74	69.14	69.93	68.91	83.74	83.03	83.39	58.45	79.50	74.74	77.04
	MTMLNet [22]	TIP'25	27.56	53.46	52.48	52.96	61.66	81.82	76.48	79.06	58.53	80.72	73.55	76.97
	DiffusionAD ★ [25]	TPAMI'25	19.50	31.72	62.71	42.13	Training Without Convergence				Training Without Convergence			
	INPFormer ★ [11]	CVPR'25	1.31	5.73	23.84	9.24	0.28	2.00	14.02	3.50	Training Without Convergence			
	TSNet [5]	ESWA'26	40.34	59.88	68.69	63.99	51.76	69.66	75.90	72.64	39.22	58.98	67.75	63.06
Multi-frame	ST-Trans [15]	TGRS'24	44.93	87.21	52.54	65.57	67.28	**93.41**	72.66	81.74	64.50	86.84	75.16	80.58
	SSTNet [1]	TGRS'24	52.25	83.49	63.25	71.98	64.87	87.17	75.09	80.68	68.21	75.70	**91.34**	82.79
	Tridos [3]	TGRS'24	49.72	88.55	56.69	69.13	73.01	87.66	84.15	85.87	61.69	83.16	74.11	78.37
	STME [13]	EAAI'25	46.57	83.51	56.19	67.18	54.06	83.79	65.22	73.35	65.11	87.09	75.52	80.89
	DTUM [7]	TNNLS'25	50.32	81.53	63.00	71.08	72.45	88.99	82.77	85.77	64.72	83.32	78.81	81.00
	ADSUNet [26]	PR'26	43.43	**92.54**	47.96	63.18	47.66	92.50	52.12	66.67	66.68	**88.76**	75.24	81.44
	CHAL (Ours)	-	**54.28**	73.08	75.26	**74.15**	**75.25**	86.35	**88.49**	**87.41**	**71.37**	86.58	83.02	**84.76**

is defined as follows:

$$\mathrm{AP} \approx \frac{1}{N}\sum_{i=1}^{N}\max_{\tilde{r}\geq r_i}\frac{\mathrm{TP}(\tilde{r})}{\mathrm{TP}(\tilde{r})+\mathrm{FP}(\tilde{r})},\ r_i=\frac{i-1}{N-1} \quad (13)$$

To ensure a monotonically non-decreasing Precision-Recall curve and smooth out fluctuations caused by varying confidence scores, we apply an all-point interpolation method to the precision function. In this way, the interpolated precision at any given recall r is defined as the maximum precision achieved at any recall $\tilde{r}$ that is greater than or equal to r. The mAP (mean Average Precision) is a comprehensive metric used to evaluate a model's overall accuracy and robustness, typically by averaging Average Precision values across multiple classes and IoU thresholds. Therefore, all mAP results reported in this paper are calculated at a fixed IoU threshold of 0.5.

E. More Experimental Results

E.1. Quantitative Comparisons

Table 2 presents more adequate quantitative comparisons on three datasets. From table, we can easily observe that our proposed CHAL consistently outperforms other detectors across various metrics. Specifically, the improvements

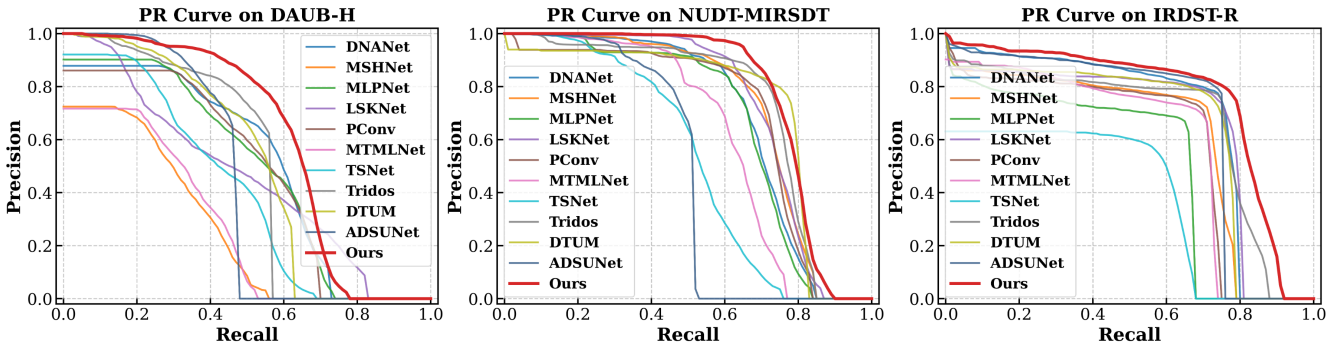

Figure 2. PR curve comparisons between different methods on DAUB-H, NUDT-MIRSDT and IRDST-R.

in mAP_{50} and F1 verify the effectiveness and superiority of our CHAL in accurately detecting moving infrared small targets within various scenes. In addition, general anomaly-detection methods cannot be directly applied to MISTD. These results further support the insight that reformulating MISTD as an anomaly discovery task is valid.

Table 3. The inference cost comparisons on DAUB-H.

Methods	Frames	mAP_{50}	F1	Param ↓	GFlops ↓	FPS ↑
ACM [2]	1	49.84	70.81	**3.04M**	**24.73**	**29.11**
ISNet [27]	1	44.73	66.46	3.49M	265.73	11.20
UIUNet [19]	1	49.23	70.47	53.06M	456.70	3.63
AGPCNet [28]	1	22.89	46.91	14.88M	366.15	4.79
DNANet [6]	1	50.76	71.52	7.22M	135.24	4.82
SIRST5K [10]	1	40.66	63.68	11.48M	182.61	7.37
MSHNet [9]	1	26.47	51.75	6.59M	69.59	18.55
MLPNet [16]	1	49.17	70.47	10.79M	34.72	5.93
SAMamba [21]	1	20.47	44.71	39.72M	274.59	3.35
PConv [23]	1	48.48	69.93	6.59M	69.29	10.01
MTMLNet [22]	1	27.56	52.96	16.79M	117.97	15.21
TSNet [5]	1	40.34	63.99	3.18M	197.81	12.18
ST-Trans [15]	5	44.93	65.57	38.13M	145.16	3.90
SSTNet [1]	5	52.25	71.98	11.95M	123.59	9.24
Tridos [3]	5	49.72	69.13	14.13M	130.72	11.23
DTUM [7]	5	50.32	71.08	9.64M	128.16	14.28
ADSUNet [26]	5	43.43	63.18	26.06M	112.84	10.52
CHAL (Ours)	5	**54.28**	**74.15**	15.69M	137.04	12.96

E.2. Inference Cost Comparisons

As shown in Table 3, we conduct more comprehensive comparisons of inference cost. From this table, we could further observe that our CHAL effectively maintains a favorable balance between detection performance and complexity. Besides, it has a medium computational cost of 15.69M and 137.04 GFlops, lower than many works, *e.g.*, the 26.06M by ADSUNet, and the 145.16 GFlops by ST-Trans. Moreover, its 12.96 FPS is higher than those of many single-frame methods, which are often regarded as approaching real deployment requirements (10 FPS) for modern high-level devices [29]. This establishes a solid foundation for deploying CHAL in real-world scenarios.

E.3. PR Curve Comparisons

To provide a more comprehensive evaluation of detection performance, we compare the PR curves of different methods on three datasets, as shown in Figure 2. From this, it is clear that our curves outperform those of the compared methods across three datasets. Specifically, on DAUB-H, our curve consistently reaches the top-right positions. This pattern continues on NUDT-MIRSDT and IRDST-R. The closer a method is to the top-right corner, the higher its validity. Therefore, these PR curves highlight the superiority of our CHAL in maintaining both high recall and high precision across diverse scenarios.

E.4. More Ablation Studies

Effects of Background Modeling Methods To further validate the effectiveness of our proposed SNF, we compare it with several representative background modeling schemes, ranging from traditional statistical methods to deep learning-based ones, as shown in Table 4. From it, we could have two evident findings. **First**, traditional statistical methods, *i.e.*, Temporal Median and Gaussian Mixture, yield the lowest performance. For instance, on DAUB-H, Temporal Median only achieves an mAP_{50} of 42.15% and an F1 of 60.34%. This is primarily because these methods rely on rigid assumptions of static scenes, making them struggle to handle the complex and dynamically evolving

Table 4. Ablation study of various background modeling schemes.

Setting	DAUB-H				NUDT-MIRSDT			
	mAP_{50}	Pr	Re	F1	mAP_{50}	Pr	Re	F1
Temporal Median	42.15	58.32	62.47	60.34	61.28	72.15	75.83	73.95
Gaussian Mixture	45.83	61.75	65.92	63.77	64.72	75.42	78.16	76.77
LSTM	48.92	66.41	68.75	67.56	68.35	79.28	82.47	80.85
3D CNN	51.47	69.83	71.62	70.71	71.86	82.64	85.19	83.90
SNF (Ours)	**54.28**	**73.08**	**75.26**	**74.15**	**75.25**	**86.35**	**88.49**	**87.41**

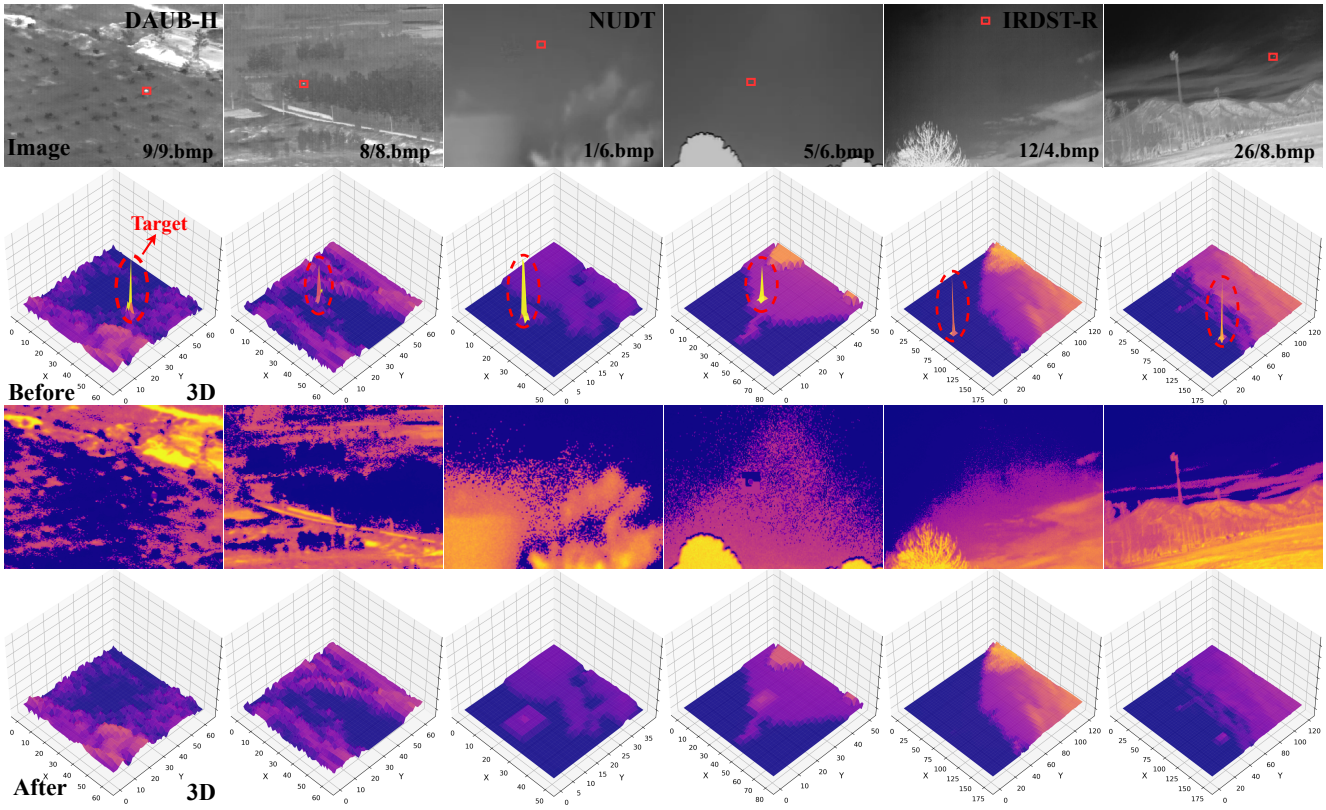

Figure 3. Visualizations of the background modeling on three datasets. Two samples for each dataset. The first and third rows show the original infrared images and predicted backgrounds $\mathcal{B}$. The second and fourth rows denote the 3D visualizations before and after SNF.

backgrounds in MISTD.

Second, deep learning-based methods, *e.g.*, LSTM and 3D CNN, significantly outperform traditional ones by learning from numerous samples. Specifically, 3D CNN serves as a strong baseline, achieving 51.47% mAP_{50} on DAUB-H. However, these methods typically operate on discrete pixel grids. They are difficult to capture the continuous spatio-temporal evolution of backgrounds, leading to suboptimal performance. In contrast, our SNF consistently achieves the best performance across all metrics on both datasets. This significant gains verifies our core motivation: modeling the background as a continuous neural field allows for a more precise reconstruction of dynamic evolution patterns.

Effects of Cosine Similarity $\mathcal{S}$ To determine the optimal metric for measuring the deviation between input frames and the predicted background in our HAL, we investigate five different measurement functions, as shown in Table 5. From table, it is obvious that cosine similarity consistently yields the best performance across all metrics on both datasets. In detail, on DAUB-H, it obtains the highest mAP_{50} of 54.28% and F1 of 74.15%. It has a substantial improvement over the second-best one, Mahalanobis distance, which achieves 52.63% mAP_{50}. This could because that cosine similarity focuses on the direction of high-dimensional feature vectors rather than their magnitude, allowing a detector to assess whether the structural pattern of a region deviates from the background normality and ignoring the interference of brightness variations. It is also worth noting that KL Divergence results in the worst performance. This is likely because deep feature maps do not strictly follow probability distributions, causing numerical instability when applying probabilistic measures directly.

Table 5. Ablation study of the cosine similarity in our HAL.

Setting	DAUB-H				NUDT-MIRSDT			
	mAP_{50}	**Pr**	**Re**	**F1**	**mAP_{50}**	**Pr**	**Re**	**F1**
Manhattan (L1)	49.85	65.42	71.38	68.28	69.74	80.15	83.67	81.87
Euclidean (L2)	51.20	67.83	72.91	70.29	71.05	82.4	85.22	83.79
Mahalanobis	52.63	70.15	73.84	71.95	72.88	84.71	86.05	85.37
KL Divergence	48.21	68.95	65.47	67.16	67.32	83.28	79.51	81.35
Cosine (Ours)	**54.28**	**73.08**	**75.26**	**74.15**	**75.25**	**86.35**	**88.49**	**87.41**

Effects of Spatio-Temporal Neural Fields Figure 3 performs additional visualization experiments before and after SNF on three datasets to visually demonstrate the effectiveness of SNF for background modeling. Specifically,

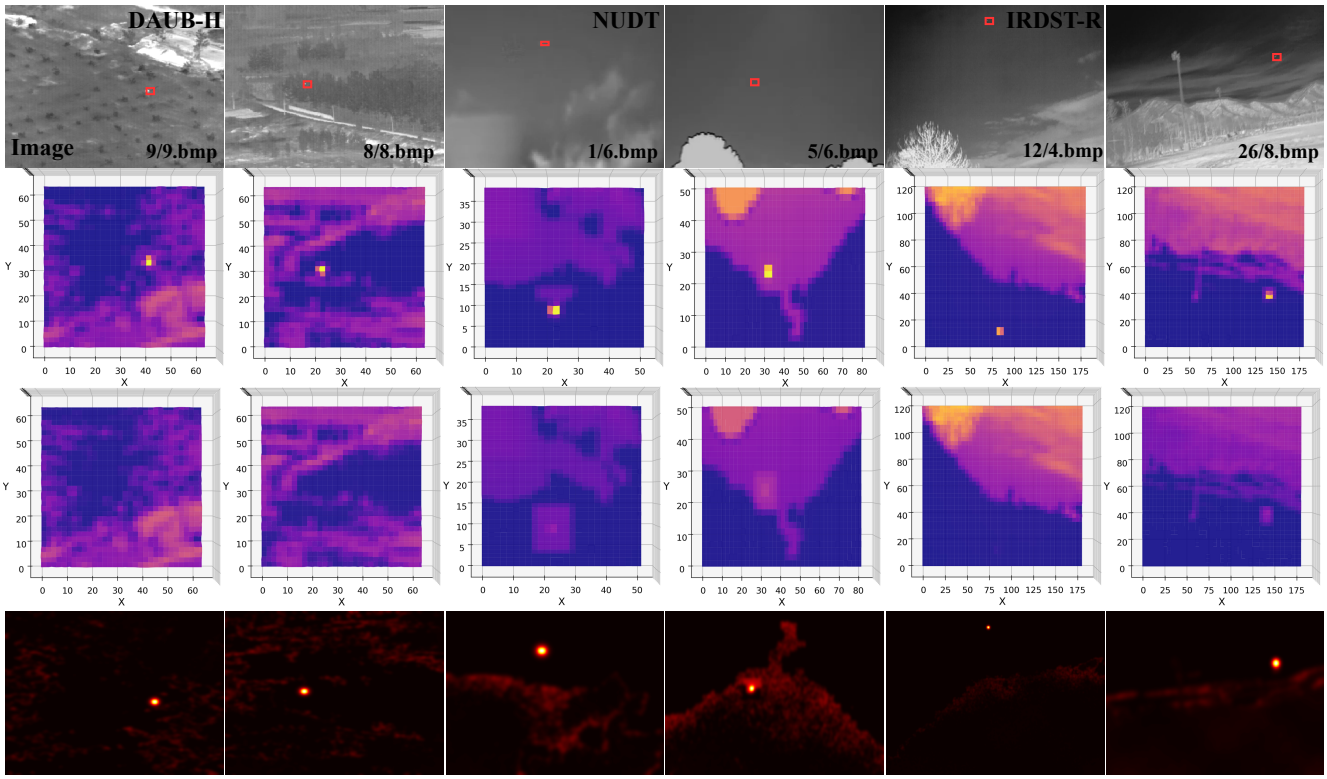

Figure 4. Visualizations of the predicted background feature $\mathcal{B}$ and appearance anomaly $\mathcal{A}_a$. The first rows are original images. The second and third rows show the 3D visualizations (top view) before and after SNF. The fourth rows denote the appearance anomaly $\mathcal{A}_a$.

we select two representative samples from each dataset. By comparison, we could have two key observations. is obvious that the features before SNF are confounded, containing both a low-frequency background plane and a peak representing anomaly signals. The other is that after employing SNF, the peak corresponding to the target appears to be removed, and it is not included in the predicted backgrounds. It clearly indicates that our SNF could successfully filter out the high-frequency target signal, providing an explicit and unbiased estimate of the background normality.

Effects of Hierarchical Anomaly-Aware Learning Figure 4 provides the crucial visualizations of the SNF predicted backgrounds and the appearance anomaly map $\mathcal{A}_a$ across various scenarios, revealing two obvious observations. **First**, it further confirms the effectiveness of our SNF. The predicted background $\mathcal{B}$ (third row) are significantly smoother than initial features, successfully isolating the low-frequency background structure while effectively removing high-frequency target signals. **Second**, while $\mathcal{A}_a$ (fourth row) clearly highlights the true target location, it is simultaneously severely contaminated by background confounders. For example, on NUDT-MIRSDT, the strong and diffuse responses from cloud edges persist in anomaly maps. It indicates that $\mathcal{A}_a$ is insufficient for reliable detection and remains susceptible to spurious correlations.

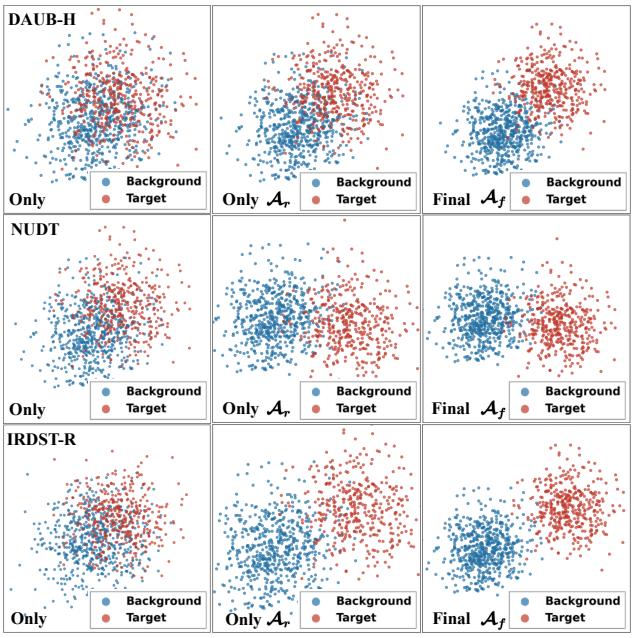

Figure 5. Feature distributions of appearance anomaly $\mathcal{A}_a$, reconstructed temporal anomaly $\mathcal{A}_r$ and final causal anomaly $\mathcal{A}_f$.

To thoroughly analyze the impacts of HAL, we select three samples from three datasets to visualize the feature

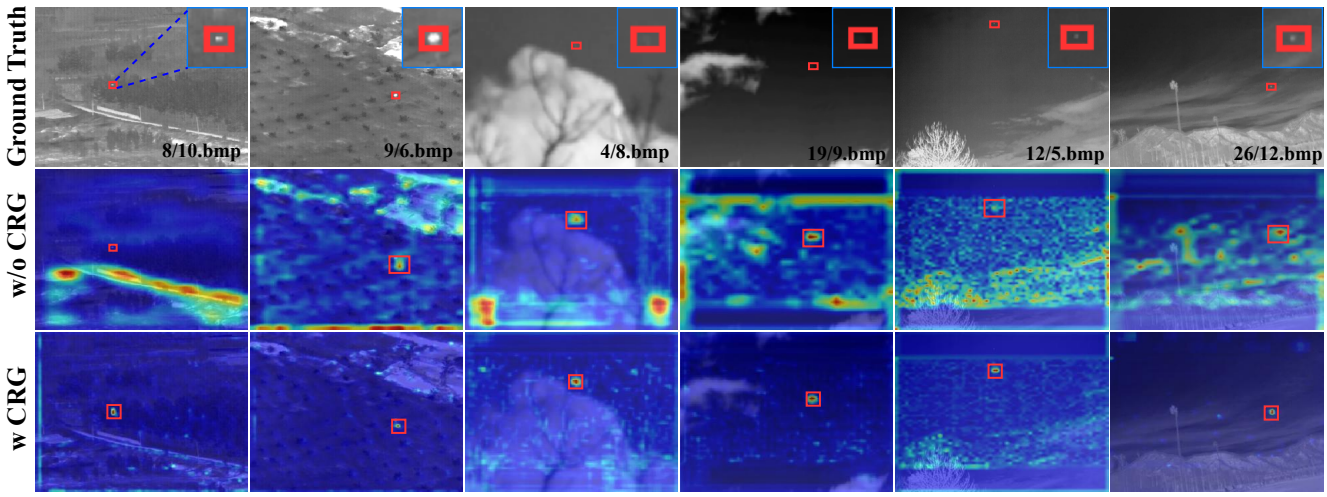

Figure 6. Feature heatmap comparisons before (w/o) and after (w) CRG. The first two columns are on DAUB-H, The middle two columns are on NUDT-MIRSDT, and the last two are on IRDST-R. The target regions (blue boxes) are amplified at the top-right corner.

distributions of targets (red) and backgrounds (blue) at different stages, as shown in Figure 5. From this, it could be easily observed that when only using the appearance anomaly $\mathcal{A}_a$, targets and backgrounds overlap significantly. When employing the reconstructed temporal anomaly $\mathcal{A}_r$, they begin to separate gradually. After applying the final causal anomaly $\mathcal{A}_f$, clear boundaries emerge. It validates our initial motivation that relying solely on appearance anomaly is insufficient to distinguish true targets from background confounders. Moreover, this clearly demonstrates the effectiveness of our HAL strategy, successfully learning a robust feature representation.

Effects of Causal Relation Guiding To visually demonstrate the effectiveness of CRG, we compare the feature heatmaps before (w/o CRG, row 2) and after (w CRG, row 3) the causal intervention, as shown in Figure 6. From figure, it is evident that the focus positions of feature heatmaps by "w/o CRG" are obscure, and targets are even lost in complex backgrounds. In contrast, after employing CRG, the feature response of small targets is notably enhanced. Meanwhile, the noisy background becomes cleaner. These results provide strong visual evidence that CRG is essential for our CHAL's high performance. It successfully disentangles the feature space, confirming that the causal intervention is superior in suppressing spurious correlations and maximizing the signal strength of true targets.

Details of Synthetic Test Set To rigorously validate the robustness of our CHAL framework against varying anomaly strengths and types, we construct several augmented test sets based on the ground truth masks of NUDT-MIRSDT. Unlike generating fully synthetic scenes, we employ a highly controlled augmentation strategy that preserves the original real targets and background clutter in the frames to maintain the realistic data distribution.

In detail, we utilize an injection strategy to add synthetic anomalies $\mathcal{T}$ onto the real infrared scenes. To ensure evaluation validity, we first identify the non-target background areas using the original ground truth masks to prevent overlap with existing real targets. The augmented frame I_{aug} is generated by

$$I_{\text{aug}}(x, y) = I_{\text{raw}}(x, y) + \mathcal{T}(x, y; \text{type}) \tag{14}$$

(a) Anomaly Strength: We quantify the anomaly strength using the Contrast-to-Noise Ratio (CNR) rather than absolute pixel intensity. It adapts the target intensity to the local background clutter, ensuring consistent difficulty levels across different scenes. For a potential injection position (x, y), we compute the local background mean intensity μ_b and standard deviation σ_b within a local surrounding window W ($e.g.$, a 10×10 annular region excluding the target center). Let $\lambda \in [0, 1]$ be the normalized anomaly strength factor. We define a maximum reference CNR, denoted as CNR_{max} (empirically set to 10.0), representing a highly salient target. By inverting the standard CNR definition, we obtain the required target mean intensity μ_t:

$$\begin{cases} \text{CNR} = \dfrac{\mu_t - \mu_b}{\sigma_b} = \lambda \cdot \text{CNR}_{\text{max}} \\ \mu_t(x, y) = \mu_b + \text{CNR} \cdot \sigma_b \end{cases} \tag{15}$$

Finally, the pixel values in the synthetic target region are set to adhere to this calculated mean intensity μ_t. It ensures that when $\lambda < 0.4$, the generated targets are submerged in the background noise (σ_b), strictly simulating infrared targets that are challenging for conventional detectors.

(b) Anomaly Type: For brightness anomaly, we model the spatial appearance using a 2D Gaussian Point Spread Function (PSF), as follows:

$$\mathcal{T}_{bright}(x) = \exp\left(-\frac{|x-\mathbf{c}_t|^2}{2\sigma^2}\right), \tag{16}$$

where $\mathbf{c}_t$ is the target center coordinates, and σ controls the thermal spread radius. Besides, we set $\sigma \in [1.0, 2.5]$ pixels to simulate targets ranging from sub-pixel point sources to slightly resolved blobs.

Real-world targets ($e.g.$, UAVs, vehicles) often exhibit non-Gaussian characteristics due to complex surface emissivity, or irregular shapes. To break the smoothness assumption, we introduce geometric irregularity and internal noise as texture anomaly. The anomaly is generated by masking a stochastic noise patch:

$$\mathcal{T}_{text}(x) = \mathcal{M}_{shape}(x) \odot \mathcal{N}(\mu_{text}, \sigma^2_{text}) \tag{17}$$

where $\odot$ denotes the Hadamard product. $\mathcal{M}_{shape}(x)$ is a binary mask generated via a Random Walk algorithm to simulate non-convex boundaries, and $\mathcal{N}$ represents pixel-wise Gaussian noise simulating internal texture variation.

In dynamic scenarios, backgrounds ($e.g.$, waving trees) often exhibit motion patterns that can be easily confused with targets. The motion anomaly is strictly defined by the kinematic independence, $i.e.$, its motion trajectory deviates from backgrounds. Specifically, we define motion anomaly in temporal domain. Let $\mathbf{u}_{bg}(x,t)$ be the local background optical flow vector estimated from adjacent frames at position x. The target's trajectory $\mathbf{c}_t$ is updated by:

$$\mathbf{c}_t = \mathbf{c}_{t-1} + \mathbf{v}_{tar}, \text{ s.t. } \|\mathbf{v}_{tar}, \mathbf{u}_{bg}(\mathbf{c}_{t-1}, t)\| < \epsilon \tag{18}$$

where $\mathbf{v}_{tar}$ is the assigned target velocity vector, ϵ is a divergence threshold as the anomaly boundary, and $\|\cdot\|$ denotes the inner product. The constraint ensures that the target's motion vector is mathematically orthogonal to or distinct from the local background motion.

Effects of Anomaly Strength and Type Figure 7 provides more adequate ablation experiments on several synthetic test sets based on NUDT-MIRSDT to explore the perception ability of our CHAL for different anomaly strengths and types. Specifically, we control the anomaly strength by adjusting the contrast between synthetic targets and local backgrounds. Three representative anomaly types are designed, $i.e.$, brightness, texture, and motion (the motion trajectory deviates from the background). From figure, we could clearly find that our CHAL notably outperforms all ablation variants. For example, when the strength is less than 0.4, the mAP_{50} and F1 of "w/o HAL" and "only $\mathcal{A}_a$" tends to be 0. In contrast, our CHAL still maintains a high performance. Moreover, this trend continues for all

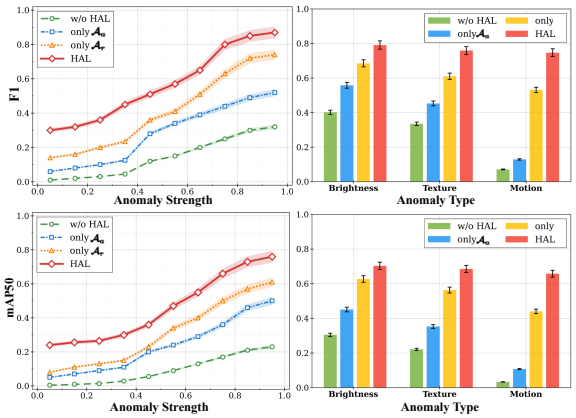

Figure 7. Ablation study of different anomaly strengths and types.

anomaly types. It shows that our CHAL could simultaneously perceive both low-level appearance anomalies and high-level motion ones, while others cannot. The experimental results further support our motivation: our CHAL could effectively address the limitations of target-centered learning by shifting to an anomaly discovery paradigm.

Effects of Hyper-Parameters **(1) Frame Number** t: To explore the effects of frame number t on detection performance, we perform a group of ablation studies, as shown in Figure 8. It is apparent that our CHAL reaches the highest mAP_{50} and F1 on both DAUB-H and NUDT-MIRSDT when $t = 5$. One possible reason is that if t is too small, the motion features of infrared small targets could not be captured effectively. Conversely, if t is too big, excessive redundant features could cause interference. Therefore, the optimal frame number of CHAL seems to be 5.

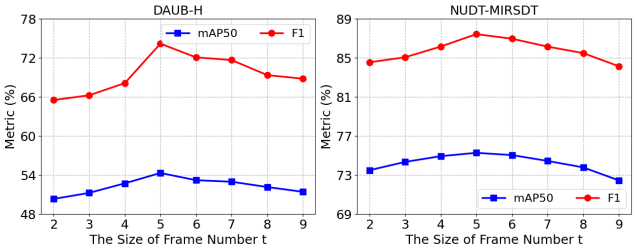

Figure 8. Effects of frame sampling number t on CHAL.

(2) The Number of Frequency Bands L: Hyperparameter L governs the spectral bandwidth of the position encoder, which is pivotal for the SNF. It determines the granularity of the spatio-temporal coordinates mapped into the high-dimensional Fourier feature space. To identify the optimal configuration, we conduct a group of ablation studies, as shown in Figure 9. From it, both two datasets exhibit a consistent ascend-peak-descend trend for mAP_{50} and F1. It is obvious that our CHAL reaches peak

performance on both DAUB-H and NUDT-MIRSDT when $L = 6$. This could because a small L implies a limited frequency bandwidth, failing to capture high-frequency background details. An excessively big L empowers the SNF to encode extremely high-frequency details, allowing the neural field to reconstruct not just backgrounds, but also the small targets themselves. As such, setting $L = 6$ provides the optimal spectral bias for MISTD.

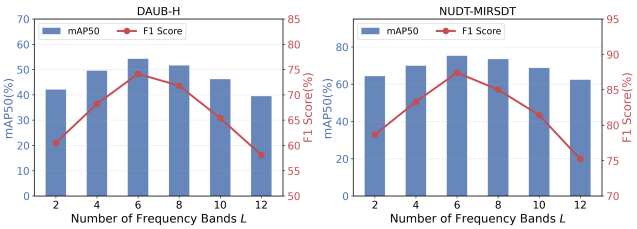

Figure 9. Effects of frequency bands number L on CHAL.

(3) Anomaly Threshold τ: The anomaly threshold τ in the CRG serves as a critical decision boundary for stratifying the causal proxy $\mathcal{A}_f$ into target-dominated and confounder-dominated strata. To determine the optimal initialization for this learnable parameter, we conducted a sensitivity analysis by varying $\tau \in [0.1, 0.9]$ on both DAUB-H and NUDT-MIRSDT, as shown in Figure 10. From figure, we could observe that the performance curves reach their peak around $\tau = 0.3$ and form a relatively flat plateau within the range $[0.2, 0.4]$. One possible reason is that if τ is too small, it allows excessive background noise and weak confounders to pass through the high-confidence stratum. Conversely, if t is larger than 0.5, it could make the detector conservative, filtering out not only confounders but also valid dim small targets. Therefore, we initialize the learnable τ to 0.3 in experiments to facilitate stable convergence.

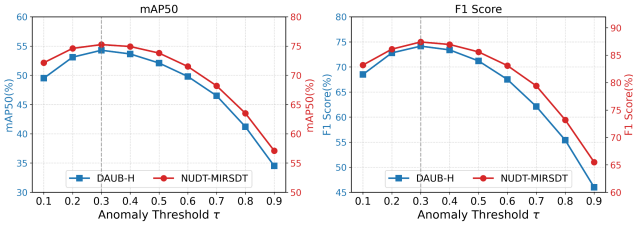

Figure 10. Effects of anomaly threshold τ on CHAL.

(4) Learnable Scale Factors η_e **and** η_s: In our CHAL framework, the enhancement factor η_e and suppression factor η_s are designed as learnable parameters to allow the detector to adaptively fine-tune the strength of causal intervention. The initialization of these parameters is decisive for convergence stability and final performance. Therefore, we perform a group of experiments to study analyze their effects, as shown in Figure 11. In it, we change η_e and η_s to different settings, while keeping other parameters fixed. From figure, it is clear that the initialization of η_e and η_e impact the detection performance at some extent. When $\eta_e = 1.4$ and $\eta_s = 0.1$, detection performance will peak on two datasets, providing the optimal inductive bias and facilitating stable convergence to the true causal effect.

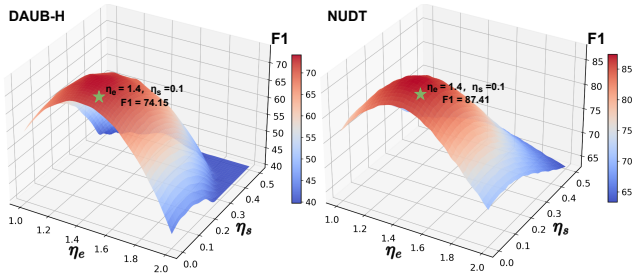

Figure 11. Ablation study of η_e and η_s on our CHAL.

F. Failure Case Analysis

To provide a holistic evaluation of our proposed CHAL and foster future research, we conduct a detailed analysis of typical failure cases observed on the test sets. As illustrated in Figure 12, although our method achieves SOTA performance, it still faces challenges in specific extreme scenarios. For example, on 14/10.bmp, it treats a bright spot as a target, causing false alarms. This could because that this distractor possesses strong visual saliency, our CRG incorrectly classifies it into the target stratum. Besides, on 15/5.bmp, a false alarm occurs at the sharp boundary of moving clouds. This could be attributed to the Spectral Bias of our SNF. With a finite Fourier bandwidth ($L = 6$), SNF struggles to perfectly reconstruct high-frequency moving edges. Moreover, on 92/1000.bmp, the target near the power lines is completely missed. Since the local background (power lines) and the target share similar high-frequency details, our SNF incorrectly encodes the target as part of the background.

These cases reveal that our anomaly discovery paradigm is relatively vulnerable when background elements mathematically resemble targets or when targets are structurally assimilated by the background. Therefore, introducing adaptive frequency modeling in SNF to decouple targets from high-frequency backgrounds and incorporating lightweight semantic logic to filter out interference are worthy of further exploration.

G. More Visualization Comparisons

For a visual comparison of the detection capabilities across various methods, we present three sets of visualizations in Figure 13-16. It is observable that our CHAL consistently demonstrates a high degree of accuracy in identifying moving infrared small targets, in contrast to compared methods,

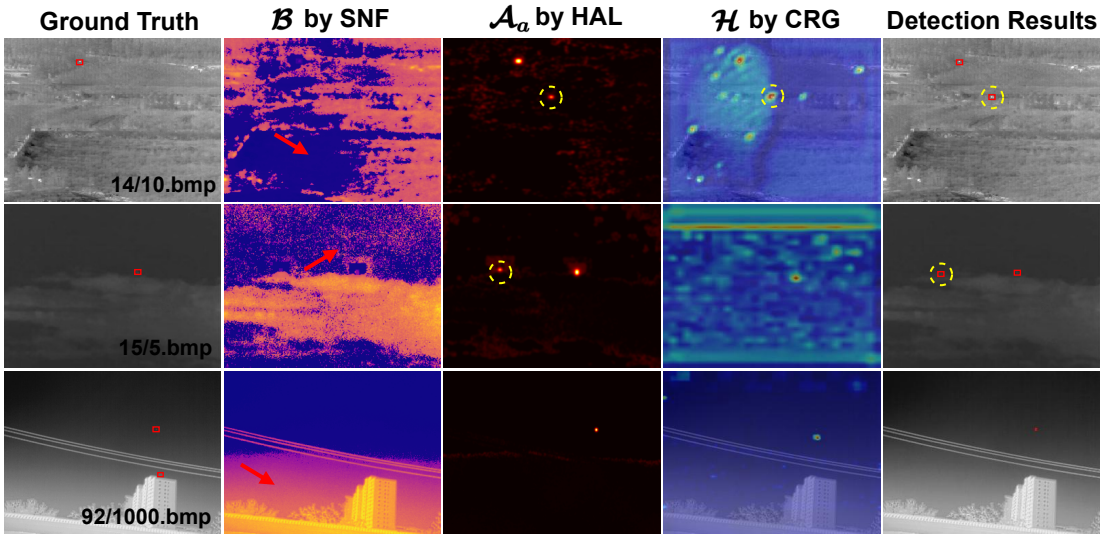

Figure 12. Failure case analysis of three representative samples on three datasets. Yellow circles are false detections.

which frequently result in a significant volume of missed and false detections. For example, in Figure 13, on the sample 10/8.bmp of DAUB-H, our CHAL could detect the correct one target, while STME and PConv DTUM detect two targets. Besides, TSNet, MTMLNet and LSKNet cannot even detect any targets, causing missed detections. In Figure 14, on NUDT-MIRSDT, our CHAL could still correctly detect the target in different cases. However, PConv, STME and SAMamba appear miss detections, *e.g.*, on the sample 4/10.bmp. Furthermore, in Figure 16, on RsCarData, our CHAL has the highest similarity with the GT on both two samples in visible-light remote sensing scenes.

Totally, these visualization comparisons above consistently support the quantitative results. It further confirms the effectiveness of our CHAL compared to other methods in detecting infrared small targets across different scenarios. Even in visible-light scenarios, it still exhibits strong adaptivity and high detection performance.

References

[1] Shengjia Chen, Luping Ji, Jiewen Zhu, Mao Ye, and Xiaoyong Yao. SSTNet: Sliced Spatio-Temporal Network With Cross-Slice ConvLSTM for Moving Infrared Dim-Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 3, 4, 5

[2] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric Contextual Modulation for Infrared Small Target Detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 949–958, 2021. 4, 5

[3] Weiwei Duan, Luping Ji, Shengjia Chen, Sicheng Zhu, and Mao Ye. Triple-domain feature learning with frequency-aware memory enhancement for moving infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 1, 3, 4, 5

[4] Bingwei Hui, Zhiyong Song, Hongqi Fan, Ping Zhong, Weidong Hu, Xiaofeng Zhang, Jianguo Lin, Hongyan Su, Wei Jin, Yongjie Zhang, and Yaxi Bai. A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background, 2019. 3

[5] Sixiang Ji, Haofei Zhang, Jingmin Zhang, Chun Fei, Xiaoyang Wang, Juanxiu Liu, and Ping Zhang. A three-stage model for infrared small target detection with spatial and semantic feature fusion. *Expert Systems with Applications*, 295:128776, 2026. 4, 5

[6] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2022. 4, 5

[7] Ruojing Li, Wei An, Chao Xiao, Boyang Li, Yingqian Wang, Miao Li, and Yulan Guo. Direction-coded temporal u-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (1):555–568, 2025. 1, 3, 4, 5

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 2

[9] Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, and Ying Fu. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition*, 2024. 4, 5

[10] Yahao Lu, Yupei Lin, Han Wu, Xiaoyu Xian, Yukai Shi, and Liang Lin. SIRST-5K: Exploring Massive Negatives Synthesis with Self-supervised Learning for Robust Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 4, 5

[11] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang,

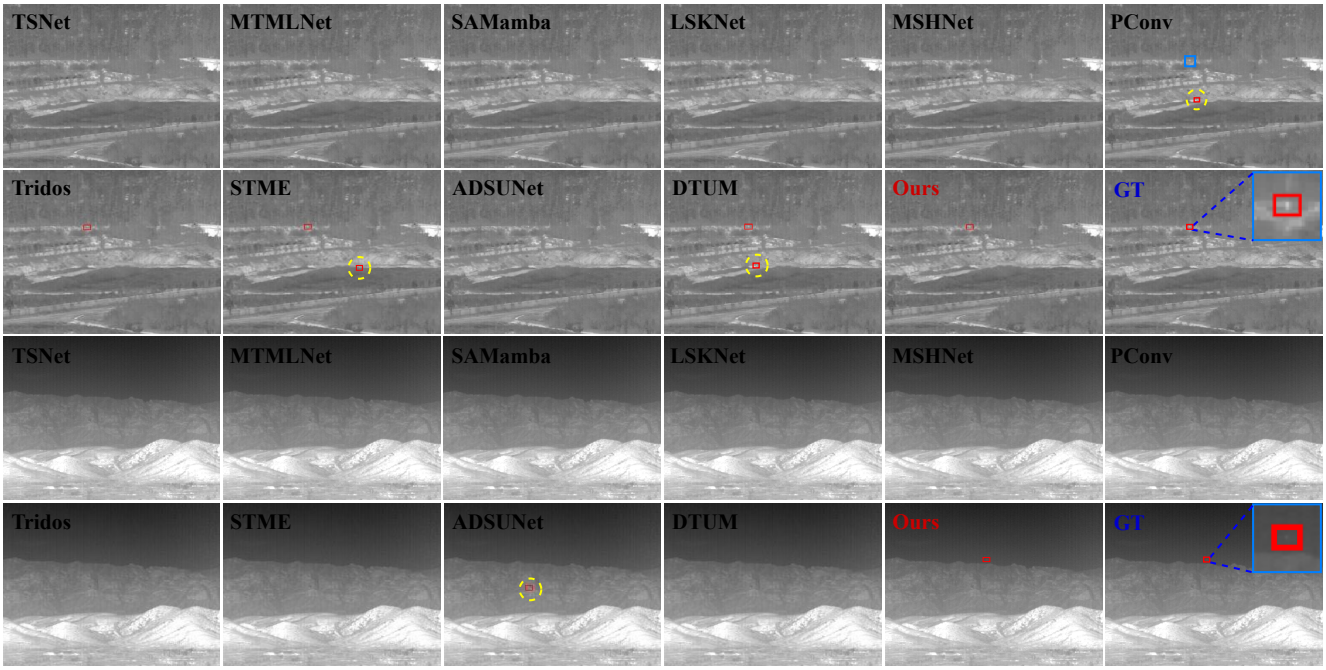

Figure 13. The visualization comparisons on the sample 10/8.bmp and 13/6.bmp of DAUB-H. GT is ground truth.

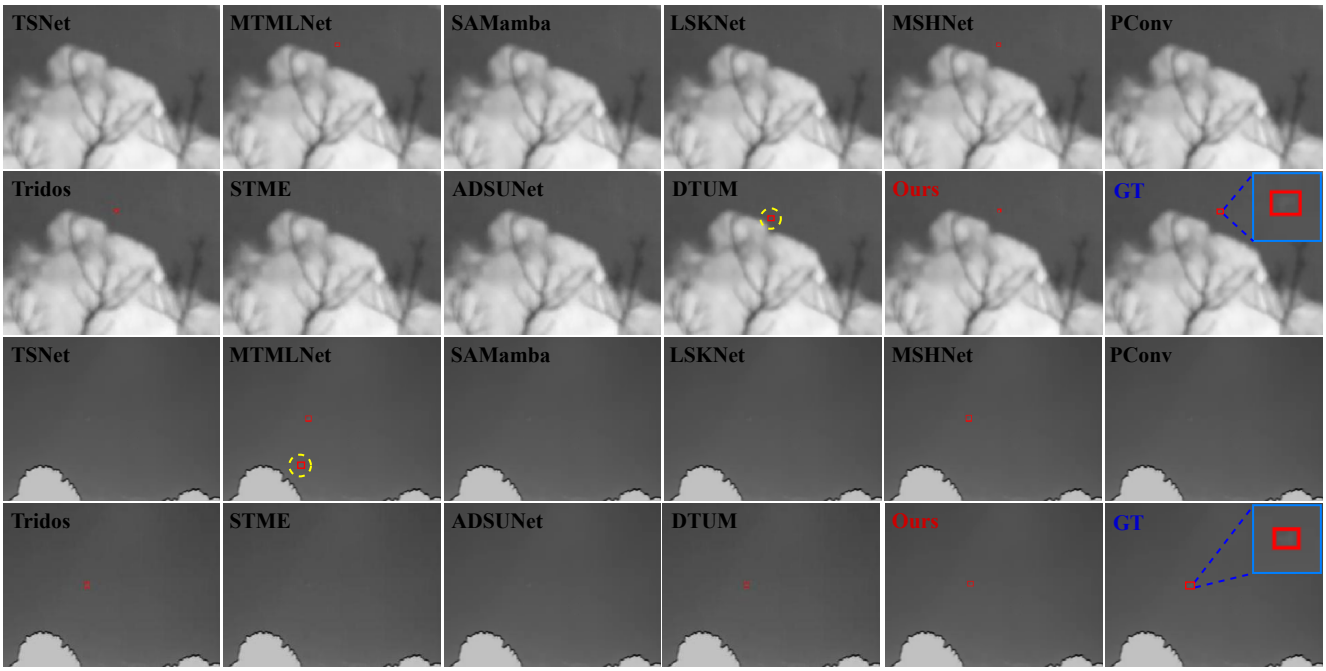

Figure 14. The visualization comparisons on the sample 4/10.bmp and 26/14.bmp of NUDT-MIRSDT. GT is ground truth.

Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 4

[12] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000. 2

[13] Shuang Peng, Luping Ji, Shengjia Chen, Weiwei Duan, and Sicheng Zhu. Moving infrared dim and small target detection by mixed spatio-temporal encoding. *Engineering Applications of Artificial Intelligence*, 144:110100, 2025. 4

[14] Heng Sun, Junxiang Bai, Fan Yang, and Xiangzhi Bai.

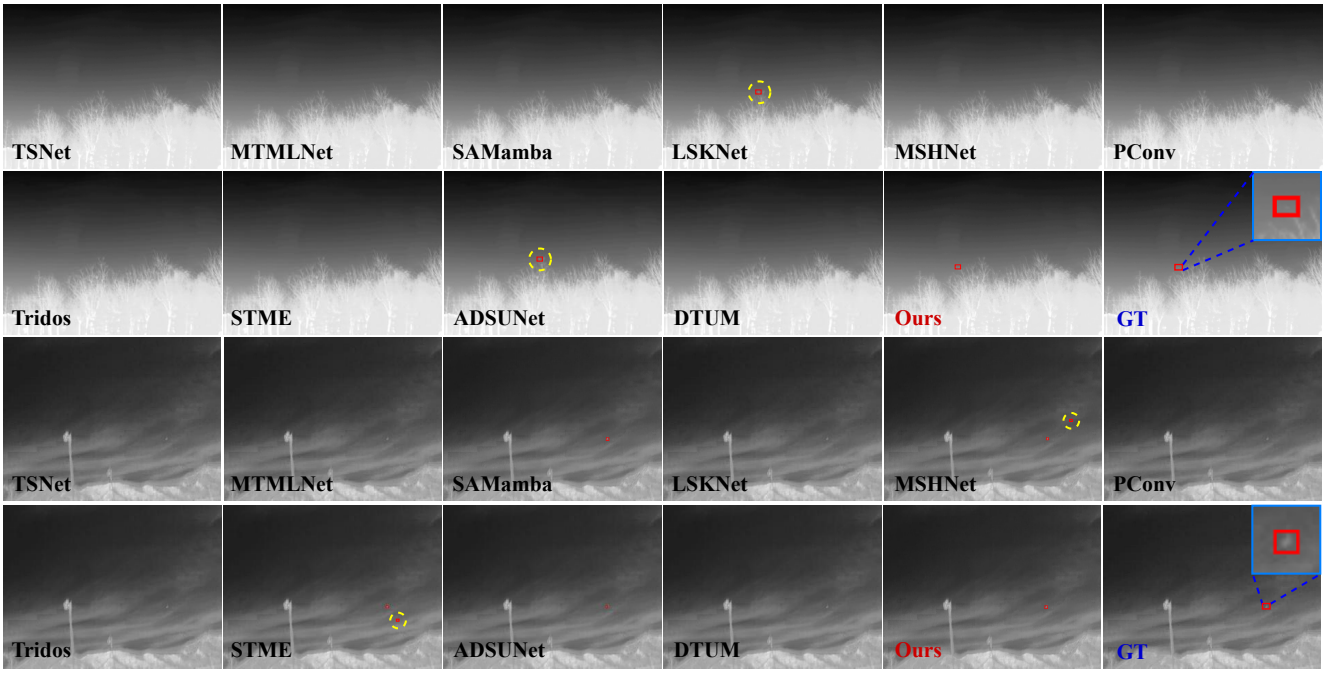

Figure 15. The visualization comparisons on the sample 4/10.bmp and 5/16.bmp of IRDST-R. GT is ground truth.

Receptive-Field and Direction Induced Attention Network for Infrared Dim Small Target Detection With a Large-Scale Dataset IRDST. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 3

[15] Xiaozhong Tong, Zhen Zuo, Shaojing Su, Junyu Wei, Xiaoyong Sun, Peng Wu, and Zongqing Zhao. ST-Trans: Spatial-Temporal Transformer for Infrared Small Target Detection in Sequential Images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 4, 5

[16] Zhishe Wang, Chunfa Wang, Xiaosong Li, Chaoqun Xia, and Jiawei Xu. MLP-Net: Multilayer Perceptron Fusion Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. 5

[17] Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, and Zhenming Peng. RPCANet: Deep Unfolding RPCA Based Infrared Small Target Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4818, 2024. 4

[18] Fengyi Wu, Anran Liu, Tianfang Zhang, Luping Zhang, Junhai Luo, and Zhenming Peng. Saliency at the helm: Steering infrared small target detection with learnable kernels. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. 4

[19] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022. 4, 5

[20] Chao Xiao, Qian Yin, Xinyi Ying, Ruojing Li, Shuanglin Wu, Miao Li, Li Liu, Wei An, and Zhijie Chen. Dsfnet: Dynamic and static fusion network for moving object detection in satellite videos. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3

[21] Wenhao Xu, Shuchen Zheng, Changwei Wang, Zherui Zhang, Chuan Ren, Rongtao Xu, and Shibiao Xu. Samamba: Adaptive state space modeling with hierarchical vision for infrared small target detection. *Information Fusion*, page 103338, 2025. 4, 5

[22] Bo Yang, Fengqian Li, Songliang Zhao, Wei Wang, Jun Luo, Huayan Pu, Mingliang Zhou, and Yangjun Pi. MTMLNet: Multi-task mutual learning network for infrared small target detection and segmentation. *IEEE Transactions on Image Processing*, 2025. 4, 5

[23] Jiangnan Yang, Shuangli Liu, Jingjun Wu, Xinyu Su, Nan Hai, and Xueli Huang. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9202–9210, 2025. 4, 5

[24] Shuai Yuan, Hanlin Qin, Xiang Yan, Naveed Akhtar, and Ajmal Mian. Sctransnet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 4

[25] Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. DiffusionAD: Norm-Guided One-Step Denoising Diffusion for Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):7140–7152, 2025. 4

[26] Liuwei Zhang, Yuyang Xi, Zhipeng Wang, Wang Zhang, Fanjiao Tan, and Qingyu Hou. ADSUNet: Accumulation-difference-based Siamese U-Net for inter-frame infrared dim and small target detection. *Pattern Recognition*, 169:111942, 2026. 4, 5

[27] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. ISNet: Shape matters for infrared

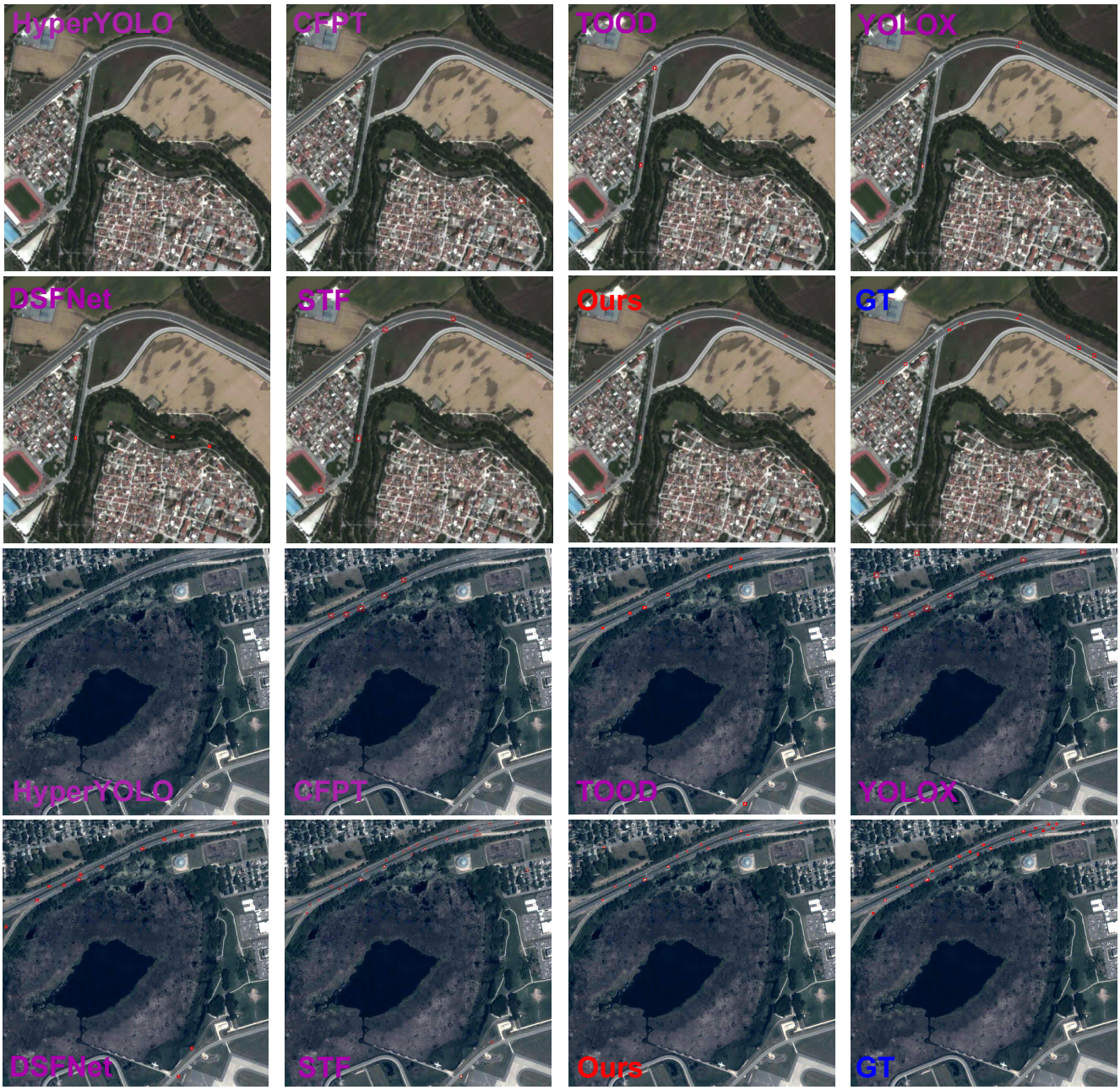

Figure 16. The visualization comparisons on the sample 6/296.jpg and 9/274.jpg of RsCarData. GT is ground truth.

small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022. 1, 4, 5

[28] Tianfang Zhang, Lei Li, Siying Cao, Tian Pu, and Zhenming Peng. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background. *IEEE Transactions on Aerospace and Electronic Systems*, 59 (4):4250–4261, 2023. 4, 5

[29] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7853–7869, 2022. 5

[30] Sicheng Zhu, Luping Ji, Jiewen Zhu, Shengjia Chen, and Weiwei Duan. TMP: Temporal Motion Perception with spatial auxiliary enhancement for moving Infrared dim-small target detection. *Expert Systems with Applications*, page 124731, 2024. 3